

[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

Google scholar

duplicate document webcrawler id OR identifier

Search

[Advanced Scholar Search](#)

[Scholar Preferences](#)

[Scholar Help](#)

Scholar All articles - **Recent articles** Results 1 - 10 of about 746 for duplicate document webcrawler id OR identifier content. (0.12 seconds)

Identifying **duplicate** documents from search results without comparing **document content**

EW Brown, JM Prager - US Patent 5,913,208, 1999 - Google Patents

... 15,1999 [54] IDENTIFYING **DUPLICATE DOCUMENTS** FROM SEARCH RESULTS WITHOUT COMPARING

DOCUMENT CONTENT [75] Inventors: Eric William Brown, New Fairfield, Conn ...

[Cited by 17](#) - [Related articles](#) - [Web Search](#) - [All 2 versions](#)

Method and system for detecting **duplicate** documents in web crawls

D Meyerzon, S Shoroff, FS Terek, S Norin - US Patent 6,547,829, 2003 - Google Patents

... The **Web crawler** program 200 may retrieve electronic **document** information for ... Detecting

Duplicate Documents Using **Content** Identifiers As mentioned above ...

[Cited by 5](#) - [Related articles](#) - [Web Search](#) - [All 2 versions](#)

[PDF] •On the evolution of clusters of near-**duplicate** web pages

D Fetterly, M Manasse, M Najork - Proceedings of the 1st Latin American Web Congress, 2003 - cwr.cl

... data using the Mercator **web crawler** [12], customized ... **documents** downloaded, were

near-**duplicates** of the 13,283,856“canonical” **documents** representing the ...

[Cited by 62](#) - [Related articles](#) - [View as HTML](#) - [Web Search](#) - [All 21 versions](#)

Understanding **Content** Reuse on the Web: Static and Dynamic Analyses- •ufmg.br (PDF)

R Baeza-Yates, A Pereira, N Ziviani - Lecture Notes in Computer Science, 2008 - Springer

... In order to design a **Web crawler**, many different aspects must be ... new **document** matches

all the previous conditions, any **duplicate** of this **document** cannot be ...

[Cited by 1](#) - [Related articles](#) - [Web Search](#) - [BL Direct](#) - [All 5 versions](#)

Managing duplicates in a web archive- •ul.pt (PDF)

D Gomes, AL Santos, MJ Silva - Proceedings of the 2006 ACM symposium on Applied computing, 2006 - portal.acm.org

... reducing the probability of an incremental **web crawler** downloading a ... on one of the

volumes, the **document** is considered to be a **duplicate** and its ...

[Cited by 12](#) - [Related articles](#) - [Web Search](#) - [All 7 versions](#)

•OverCite: A distributed, cooperative CiteSeer

J Stribling, J Li, IG Council, MF Kaashoek, R ... - Proc. 2006 NSDI, 2006 - usenix.org
... 3.4 **Web Crawler**. ... eg, title, authors, citations, etc.) as well as the bare ASCII text
of the **document** - and checks whether this is a **duplicate document**. ...

[Cited by 12](#) - [Related articles](#) - [Cached](#) - [Web Search](#) - [All 28 versions](#)

Understanding **Content** Reuse on the Web: Static and Dynamic Analyses

S Barcelona - Advances in Web Mining and Web Usage Analysis: 8th ..., 2007 - books.google.com
... In order to design a **Web crawler**, many different aspects must be ... new **document** matches
all the previous conditions, any **duplicate** of this **document** cannot be ...

[Related articles](#) - [Web Search](#)

System and method for classifying electronically posted documents

AWL Huang, N Sundaresan - US Patent App. 11/526,470, 2006 - Google Patents
... search engine typically uses proprietary **webcrawler** and indexing ... to be **duplicates**,
the **duplicate** meta- data ... module reads the downloaded **document** and generates ...

[Related articles](#) - [Web Search](#) - [All 4 versions](#)

OverCite: A cooperative digital research library- •psu.edu (PDF)

J Stribling, IG Council, J Li, MF Kaashoek, DR ... - Lecture notes in computer science, 2005 - Springer
... peer-to-peer search optimizations [23,4]. **Web crawler**. ... If the **document** is not a
duplicate, the crawler ... PC which papers to read (using the **document**-alert feature ...

[Cited by 47](#) - [Related articles](#) - [Web Search](#) - [Library Search](#) - [BL Direct](#) - [All 38 versions](#)

Merging techniques for performing data fusion on the web- •tugraz.at (PDF)

T Tsirikas, M Laimas - Proceedings of the tenth international conference on ..., 2001 - portal.acm.org
... com) and **Webcrawler** (<http://www.webcrawler.com>), selected ... set to 30, since the more
documents retrieved, the ... an increase on the number of **duplicate documents** is ...

[Cited by 15](#) - [Related articles](#) - [Web Search](#) - [All 6 versions](#)

Key authors: [M Najork](#) - [D Fetterly](#) - [J Stribling](#) - [M Manasse](#) - [M Kaashoek](#)

Goooooooooooooogle ►

Result Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#)

duplicate document webcrawler id OR identifier content

Search

[Go to Google Home](#) - [About Google](#) - [About Google Scholar](#)

©2009 Google